BUSS-2503 Regression Analysis

Final Project

Due Date: May 27, 2025, 11:00 am

《灰龟》公司：体检数据助力癌症预测

《灰龟粉犀》公司 (以下简称《灰龟》公司) 是一家专注于医疗数据分析的创新型公司，致力于通过数据洞察和技术创新为健康管理提供更精准的解决方案。我们想要通过分析体检数据，帮助大家提前发现结直肠癌的潜在风险，提供一份“健康警报”单，为高风险群体提供及时干预。为了完成这一伟大的任务，灰龟公司招募了一支数据分析团队，也就是回归分析课上的同学们，来设计分类模型，通过体检指标来预测是否患上结直肠癌。

《灰龟》公司有如下需求：

- 预测目标：使用体检指标预测结直肠癌的发生。模型的评估指标为 **F1 分数**。
- 对数据进行清洗与预处理，包括处理缺失值、异常值等等。
- 《灰龟》公司对回归模型情有独钟，请**不要**使用课程范围外的其他机器学习或深度学习模型。阐明你在分析中所作的假设，建立回归模型进行预测和分析。
- 请对数据的处理充满想象力，这个会是最终取得更好效果的关键。
- 完成一个不超过 6 页的报告，并进行一个 5 分半钟的演讲。

数据说明:

- y: 表示病人是否有癌变，1 为有，0 为无
- $X_1 - X_5$: 钾，钠，氯，葡萄糖，葡萄糖,
- $X_6 - X_{10}$: 肌酐，尿素，癌胚抗原，糖类抗原 CA199，总胆红素,
- $X_{11} - X_{15}$: 直接胆红素，间接胆红素，总蛋白，白蛋白，球蛋白,
- $X_{16} - X_{20}$: 白蛋白/球蛋白，丙氨酸氨基转移酶，门冬氨酸氨基转移酶，碱性磷酸酶，总胆汁酸,
- $X_{21} - X_{25}$: -谷氨酰转肽酶，前白蛋白，乳酸脱氢酶，二氧化碳结合力，尿酸,
- $X_{26} - X_{30}$: 白细胞计数，淋巴细胞计数，单核细胞计数，中性粒细胞计数，嗜酸粒细胞计数,
- $X_{31} - X_{35}$: 嗜碱粒细胞计数，红细胞计数，血红蛋白，红细胞比积，平均红细胞体积,
- $X_{36} - X_{40}$: 平均红细胞血红蛋白含量，平均红细胞血红蛋白浓度，红细胞分布宽度，血小板计数，平均血小板体积,
- $X_{41} - X_{45}$: 血小板压积，血小板分布宽度，大血小板细胞比率，粪隐血，粪转铁蛋白。

(变量 x 中有些变量为 0,1 变量，其中 0 为正常，1 为异常; 该数据也存在部分缺失值)

注意事项：

1. 如果你有自己想要研究的数据集和问题，并可以使用课上所学习的方法，可以在与唐老师讨论并得到同意后替换以上项目。
2. 数据以及提交预测结果的地方在 Kaggle 系统中。**请使用如下链接进入**。

https://www.kaggle.com/t/6d9d5f68ea86472b8f5b489ddd6161fc

3. 大作业为一人一组，需要在 Kaggle 上注册你的 Team Name(不需要使用真实姓名)，但是需要将你的 Team Name 与学号和姓名在注册完成后通过以下问卷告知任课老师与助教，方便进行最终评分。请不晚于 **4 月 30 日**填写如下问卷。

   https://www.wjx.cn/vm/edJ4k2F.aspx

4. 所提交的结果应与 sample_solution.csv 的文件格式相同，其中在 y 的那一列放上你对于 test.csv 中数据的预测值。
5. Kaggle 每天提交的上限为 30 次，因此不要拖到最后时刻再上手，可能会因为次数不够无法获得较好地分数。Kaggle 上会实时显示每个账号的当前分数，但是这个分数是基于公开测试集（测试集的 60%）。最终项目结束后还会根据私有测试集（测试集的 40%）进行一个评价。
6. **不可以在 Kaggle 上提交使用本课程大纲之外的高阶机器学习算法得到的预测结果**。(如不确定，可询问唐老师确认方法是否可用)。
7. 最终提交在 Canvas 上的内容:
   - 一个大作业报告，pdf 格式。总长度**不超过 6 页**。
   - 一个 ppt 演示文档。我们会直接下载你所提交的演示文档，在最后两周课上进行汇报演讲，汇报时间**不超过 5 分半钟**。汇报演讲无着装要求。ppt 演示文档与现场演讲中英文皆可。(报告的时间与顺序按照后续 Canvas 上发布的信息为主。)
   - 一个程序文件，该文件需包含 Kaggle 上所获得结果最好的方法。
8. 重要截止日期:
   - **4 月 30 日晚上 11 点 59 分**，问卷提交 Team Name
   - **5 月 27 日上午 11 点 00 分**，Kaggle 截止
   - **5 月 27 日上午 11 点 00 分**，在 canvas 上提交电子 pdf 报告、ppt 演示文档、以及程序文件。
   - **5 月 27 日上课前**，在教室提交纸质版报告，并附上姓名与学号。
9. 评分标准:
   - 测试集上的预测精度: 10% (其中公开测试集占 60%, 私有测试集占 40%)
   - 文字报告：60%
   - 报告 PPT 与演讲: 30%
10. 你的报告中可以包括但不限于对以下问题的思考：
    - 除了最终的模型之外，你还做了哪些尝试，与你最终选择的模型相比有什么优劣?
    - 哪些变量或者他们之间的共同作用对结直肠癌产生较大的影响？你是如何得出该判断的。
    - 是否存在冗余或对模型贡献很小的特征？移除或组合这些特征会如何影响模型性能?
    - 在预测癌症这种医学问题中，假阳性和假阴性的代价各是什么？我们如何选择合适的阈值或性能指标
    - 模型是否出现过拟合或欠拟合现象？如何有效地评估和改善模型泛化能力?
    - 对于这个数据集有什么其他的发现，以及可以对《灰龟》公司在使用模型时有哪些的建议与意见?
    - 如何设计未来的数据收集策略，以提升模型的预测准确性?
    - 假如你作为医生的助手，来了两位新的患者，分别为 test.csv 的第 6 和第 7 位患者 (X_1 的值为 4.2 的两位)，你对他们的诊断如何，以及给他们什么样的建议?

Grey Turtle Inc.: Using Physical Examination Data to Aid Cancer Prediction

Grey Turtle Pink Rhino Inc. (hereafter referred to as ”Grey Turtle”) is an innovative company focused on medical data analysis. It is dedicated to providing more precise health management solutions through data insights and technological innovation. Our goal is to analyze physical examination data to help identify the potential risk of colorectal cancer early on and provide a “health alert” report for timely intervention among high-risk individuals.

To achieve this mission, Grey Turtle has recruited a data analysis team —that is, the students of this regression analysis course —to design a classification model to predict the likelihood of colorectal cancer based on physical examination indicators.

Grey Turtle Inc. has the following requirements:

- Prediction Goal: Use health check indicators to predict the occurrence of colorectal cancer. The evaluation metric is the **F1 score**.
- Clean and preprocess the data, including handling missing values, outliers, etc.
- Grey Turtle favors regression models. Please **do not** use machine learning or deep learning models beyond the course syllabus. Clearly state the assumptions made in your analysis, and build regression models for prediction and analysis.
- Be imaginative in how you handle the data —this will be the key to achieving better results.
- Prepare a report not exceeding 6 pages, and deliver a 5.5-minute presentation.

Data Description:

- y: Indicates whether the patient has cancer (1 = yes, 0 = no)
- $X_1 - X_5$: Potassium, Sodium, Chloride, Glucose, Glucose (yes, twice)
- $X_6 - X_{10}$: Creatinine, Urea, CEA (carcinoembryonic antigen), CA199 (carbohydrate antigen), Total Bilirubin
- $X_{11} - X_{15}$: Direct Bilirubin, Indirect Bilirubin, Total Protein, Albumin, Globulin
- $X_{16} - X_{20}$: Albumin/Globulin Ratio, ALT (Alanine Transaminase), AST (Aspartate Transaminase), ALP (Alkaline Phosphatase), Total Bile Acid
- $X_{21} - X_{25}$: GGT (Gamma-glutamyl transferase), Prealbumin, LDH (Lactate Dehydrogenase), CO2-binding capacity, Uric Acid
- $X_{26} - X_{30}$: WBC, Lymphocyte Count, Monocyte Count, Neutrophil Count, Eosinophil Count
- $X_{31} - X_{35}$: Basophil Count, RBC, Hemoglobin, Hematocrit, MCV (Mean Corpuscular Volume)
- $X_{36} - X_{40}$: MCH, MCHC, RDW (Red cell distribution width), Platelet Count, MPV (Mean Platelet Volume)
- $X_{41} - X_{45}$: PCT (Plateletcrit), PDW (Platelet Distribution Width), Large Platelet Ratio, Fecal Occult Blood, Fecal Transferrin

(Some variables are binary indicators, where 0 = normal, 1 = abnormal; some values are missing.)

Important Notes:

1. If you have your own dataset and research problem and can apply the methods learned in class, you may substitute the above project with prior approval from Dr.Tang.
2. Data and submission portal are on Kaggle. **Please use the following link to access:**

   https://www.kaggle.com/t/6d9d5f68ea86472b8f5b489ddd6161fc

3. This is an individual project. You must register your Team Name on Kaggle (do not use your real name), and submit your Team Name along with your student ID and name via the following form by **April 30**:

   https://www.wjx.cn/vm/edJ4k2F.aspx

4. Your submitted results should match the format of samplesolution.csv, with the predicted values for y corresponding to the rows in test.csv.
5. Kaggle limits submissions to 30 per day. Don' t leave your work to the last minute —you may not have enough submissions left to get a good score. The real-time leaderboard on Kaggle reflects performance on the public test set (60%), while final evaluation includes the private test set (40%).
6. **Please do not submit results obtained using high-level machine learning algorithms beyond the course syllabus.** (If in doubt, ask Dr.Tang.)
7. Final submission on Canvas should include:
   - A written report (PDF, **max 6 pages**).
   - A PowerPoint presentation file. This will be used for in-class presentations in the final two weeks. The presentation should be **no more than 5.5 minutes** long. No dress code. You may use either English or Chinese in the slides and speech.
   - A program file containing the code that produced your best Kaggle result.
8. Key Deadlines:
   - **April 30, 11:59 PM**: Submit Team Name via the questionnaire
   - **May 27, 11:00 AM**: Kaggle submission closes
   - **May 27, 11:00 AM**: Submit PDF report, PPT slides, and program file to Canvas
   - **Before class on May 27**: Submit a printed copy of the report in class, with your name and student ID
9. Grading Criteria:
   - Prediction Accuracy on Test Set: 10% (Public Test Set 60%, Private Test Set 40%)
   - Written Report: 60%
   - Slides and Presentation: 30%
10. You are encouraged (but not limited) to discuss the following questions in your report:
    - Aside from your final model, what other models or approaches did you try? How do they compare?

- Which variables or interactions seem to have the strongest influence on colorectal cancer? How did you determine this?
- Are there redundant features or ones with low predictive power? How does removing or combining them affect performance?
- For this medical prediction task, what are the consequences of false positives and false negatives? How can we choose an appropriate threshold or evaluation metric?
- Did you observe signs of overfitting or underfitting? How did you evaluate and improve generalization?
- What other insights did you discover from the dataset, and what advice would you give Grey Turtle on model usage?
- How would you design future data collection to improve model accuracy?
- Imagine you're a doctor' s assistant. Two new patients arrive —patients 6 and 7 in test.csv, both with $X_1 = 4.2$. What's your diagnosis and what advice would you give them?